taxaTarget: Fast, sensitive, and precise classification of microeukaryotes in metagenomic data

Seth Commichaux (✉ Seth.Commmichaux@fda.hhs.gov)

Center for Food Safety and Applied Nutrition, Food and Drug Administration, Laurel, MD

Kiran Javkar

Department of Computer Science, University of Maryland, College Park, MD

Harihara Subrahmaniam Muralidharan

Department of Computer Science, University of Maryland, College Park, MD

Padmini Ramachandran

Center for Food Safety and Applied Nutrition, Food and Drug Administration, College Park, MD

Andrea Ottesen

Center for Food Safety and Applied Nutrition, Food and Drug Administration, Laurel, MD

Hugh Rand

Center for Food Safety and Applied Nutrition, Food and Drug Administration, College Park, MD

Mihai Pop

Department of Computer Science, University of Maryland, College Park, MD

Research Article

Keywords: Eukaryotes, eukaryome, metagenomics, supervised learning, taxonomic classification

Posted Date: March 11th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1186624/v3>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

Microbial eukaryotes are nearly ubiquitous in microbiomes on Earth and contribute to many integral ecological functions. Metagenomics is a proven tool for studying the microbial diversity, functions, and ecology of microbiomes, but has been underutilized for microeukaryotes due to the computational challenges they present. For taxonomic classification, the use of a eukaryotic marker gene database can improve the computational efficiency, precision and sensitivity. However, tools which use marker gene databases often implement universal thresholds for classification rather than dynamically learning the thresholds from the database structure, impacting the accuracy of the classification process.

Results

Here we introduce taxaTarget, a method for the taxonomic classification of microeukaryotes in metagenomic data. Using a database of eukaryotic marker genes and a supervised learning approach for training, we learned the discriminatory power and classification thresholds for each 20 amino acid region of each marker gene in our database. This approach provided improved sensitivity and precision compared to other approaches, with rapid runtimes and low memory usage. Additionally, taxaTarget was better able to detect the presence of multiple closely related species as well as species with no representative sequences in the database. Several algorithms were implemented, including threshold padding, which effectively handled the scarce training data available and reduced classification errors. Using taxaTarget on metagenomes from human fecal microbiomes, a broader range of genera were detected, including multiple parasites that the other tested tools missed.

Conclusion

Data-driven methods for learning classification thresholds from the structure of an input database can provide granular information about the discriminatory power of the sequences and improve the sensitivity and precision of classification. These methods will help facilitate a more comprehensive analysis of metagenomic data and expand our knowledge about the diverse eukaryotes in microbial communities.

Background

Eukaryotes, bacteria, and archaea comprise the domains of life on Earth (according to our current understanding). Well known eukaryotes, like plants and animals, belong to two small branches on a much larger phylogenetic tree of mostly uncharacterized microscopic eukaryotes, i.e., microeukaryotes. Some microeukaryotes are well known because they benefit humans. For example, *Saccharomyces cerevisiae* (baker's yeast) and *Aspergillus sojae* are involved in the creation of fermented foods, drinks, and spices such as beer, bread, miso, and soy sauce. Others, like *Plasmodium falciparum*, *Trypanosoma brucei*, and *Cyclospora cayentanensis* are infamous for causing human diseases like malaria, sleeping sickness and food poisoning outbreaks, respectively. However, studies estimate that less than 5% of the

predicted millions of species of microeukaryotes are currently characterized [1-3]. Furthermore, the eukaryotic tree of life is still expanding as whole new supergroups are being discovered [4].

Microeukaryotes do not exist in isolation, rather they are ecologically integral members of diverse microbial communities (alongside bacteria, archaea, viruses) i.e., microbiomes, found nearly everywhere on Earth. For example, phytoplankton, a mixture of photosynthetic microorganisms floating in oceans and freshwater bodies, are responsible for nearly half of all carbon fixation and photosynthetic primary production on Earth—70% is performed by microeukaryotic groups i.e., Diatoms, Coccolithophores, and Chlorophytes [5]. In the human gut microbiome, some species (e.g., *Blastocystis* and *Entamoeba*), formerly considered strict parasites, might promote healthy and diverse assemblages of bacterial species and only express pathogenicity in certain microbial contexts [6-11].

A useful tool for exploring the diversity of microeukaryotes is metagenomics. Metagenomics is the sequencing of DNA fragments, referred to as reads, extracted directly from a sample. Such data provides a means for studying the diversity, functional potential, and ecology of microbiomes. Although metagenomics has led to an explosion of knowledge about microbial communities [12], most studies have focused on bacteria, with relatively few (although a growing number) focusing on microeukaryotic groups such as fungi and protists [9, 13-15].

The metagenomic study of microeukaryotes has lagged for a variety of reasons. Perhaps most important, is that there is a general sparsity of sequenced genomes and transcriptomes, with more eukaryotic species represented in protein databases, e.g., ~40,000 species in the UniRef100 database, than genome databases, e.g., ~4,000 species in the National Center for Biotechnology information (NCBI) Assembly database. Combined, this is still orders of magnitude fewer species than the millions estimated to exist [1, 2]. Sparse reference sequences can negatively impact the performance of tools that use sequence similarity for taxonomic classification. For example, metagenomic reads from a novel species might not align to any sequences in the reference database of a tool and thus go undetected.

Another challenge for metagenomics is that microeukaryotes often occur at low abundance in microbial communities with genomes that are orders of magnitude larger than bacteria [16, 17]. Ideally, all the genomes in a metagenome could be directly reconstructed (i.e., assembled) from the reads. However, the limitations of current sequencing technologies, and the general sparsity of eukaryotic reads per metagenome present algorithmic challenges for assembly that often result in few or no overlapping reads to construct longer eukaryotic sequences (i.e., contigs). Microeukaryotic contigs longer than 3,000 bp can sometimes be assembled, given sufficient depth of coverage [18], and classified as eukaryotic based upon k-mer composition with high accuracy [19]. However, current bioinformatic approaches for characterizing microeukaryotes tend to rely upon metagenomic reads for analysis.

An alternative to the classification of assembled contigs is mapping the reads to a reference database of sequences (e.g., genomes or genes). However, for large eukaryotic genomes (sometimes billions of base pairs), mapping metagenomic reads directly can be computationally expensive in terms of runtime and memory. Additionally, the assemblies of eukaryotic genomes in reference databases are frequently

contaminated with bacteria, potentially leading to classification errors [20-22]. A more robust and scalable solution is the use of marker sequence databases which can be clade-specific (as used in MetaPhlan3 [23]) or universally present such as ribosomal subunits (as used by Metaxa2 [24] and mTAGs [25]) or single copy orthologs (as used by EukDetect [22, 26]). The latter have shown high specificity for eukaryotes with a strong phylogenetic signal [18, 26]. A drawback for using a database of marker sequences is that the combined length of the marker sequences is usually a small fraction of any eukaryotic genome, potentially resulting in low sensitivity when employed in metagenomic data.

A general challenge for tools using sequence similarity for taxonomic classification is that sequences, and regions of sequences, can have variable levels of discriminatory power because they evolve at different rates [27]. For example, regions of the 18S ribosomal subunit (rRNA), a eukaryotic marker sequence used by Metaxa2 and mTAGs for taxonomic classification, are known to lack discriminatory power at the species, genus, and sometimes even the family or broader levels for some eukaryotic groups [3]. One consequence is that a metagenomic read can align equally well to a sequence in multiple species, providing insufficient evidence for classification at the species level. EukDetect addressed this problem by removing marker genes from its database to which bacterial reads mapped, or if the marker gene Hidden Markov Models (HMM) misclassified bacterial sequences as eukaryotic [22]. This approach, which was not extended to other groups like archaea or viruses, reduced the number of marker gene families used for taxonomic classification by EukDetect from 255 to 214. One drawback of discarding entire eukaryotic marker gene families with high similarity regions in bacteria is that unshared regions of the marker genes, which might be informative for eukaryotic taxonomic classification, also get discarded.

Another issue for approaches using sequence similarity for taxonomic classification is the application of universal alignment score cutoffs for classifying metagenomic reads that map to database sequences. For example, reads might be classified as the species of the database sequences they align to if the alignment is $\geq 95\%$ identity. This could be extended to genus if the alignment is $\geq 85\%$ identity, family if the alignment is $\geq 75\%$ identity, and so on. EukDetect and Metaxa2 use similar strategies [22, 24], but such methods ignore that sequences can evolve at different rates and can have variable discriminatory power. A more sophisticated approach, used by CCMetagen, learned the alignment score cutoffs, from phylum to species, using the sequence similarity and taxonomic relatedness of fungal marker sequences [28]. Although a data-driven approach, the thresholds learned for fungal sequences end up being applied to the entire NCBI Nucleotide database which contains sequences across the tree of life and viruses. Another approach for the taxonomic classification of metagenomic reads, used by mTAGs [25], is to assign the lowest common ancestor of the database sequences each read aligns to. However, such approaches tend to classify reads with too high of specificity when using sparse databases (such as exists for microeukaryotes) and to be too conservative when using highly representative databases [29, 30].

As an alternative, supervised learning approaches have been developed for dynamically determining thresholds using marker genes [31] and for the discovery of protein families [32] in metagenomic data; however, these approaches were developed for bacteria using bacterial marker genes or for functional

profiling instead of taxonomic classification. Supervised learning methods use labeled training data to learn functions for classifying new examples. For example, MetaPhyler [31] predicts the taxonomic relatedness (from phylum to species) of metagenomic reads aligned to marker genes. It learns the cutoffs for classification thresholds by training on the scores of reads, from the genomes of known taxa, aligned to the marker genes in its database. This approach automatically assesses if sequences contain enough information for classification at different taxonomic levels. However, a limitation of MetaPhyler is that it builds classifiers for whole genes, treating all regions of a gene as having equal classification power. ROcker [32] introduced a method for assessing if metagenomic reads were sequenced from specific gene families using sliding window, region-specific classifiers. Although a functional profiling tool, we thought the approach could be successfully adapted for taxonomic classification.

Here, we present a metagenomic taxonomic profiling tool, taxaTarget, specifically designed for microeukaryotes. We show that taxaTarget is computationally efficient, with higher sensitivity, and often higher precision than similar tools using marker sequence databases—Metaxa2, mTAGs, and EukDetect—even for species not represented in the database. Using a similar database of single copy orthologs as EukDetect, we show how a supervised learning approach automates the assessment of which genes and regions are most informative for classification, reducing the need for manual curation, and retaining more sequences in the database. Further, the supervised learning approach of taxaTarget replaces universal classification thresholds (i.e., universally applied to all sequences in a database) with classifiers tuned by sequence region in a data-driven manner. We also describe several novel solutions for effectively handling the scarcity of training data for microeukaryotes which initially led to many classification errors.

Methods/materials

Definitions

Precision = True Positives / (True Positives + False Positives)

Sensitivity = True Positives / (True Positives + False Negatives)

False Positive Rate = False Positives / (False Negatives + True Negatives)

Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)

Tools selected for benchmarking taxaTarget

To benchmark the performance of taxaTarget we selected several existing tools (Metaxa2, mTAGs, and EukDetect) that were representative of the main marker gene approaches for the taxonomic classification of microeukaryotes in metagenomic data. These approaches can be broadly classified into three categories based upon the marker sequences they use as a database. The first approach uses a database of markers that were originally used as barcodes for amplicon-based metagenomics such as the COI

gene and the ribosomal subunits. Here we selected Metaxa2 and the recently published mTAGs. The second approach uses a database of clade-specific marker genes. The only tool we could find that used this approach was MetaPhlan3, which we decided not to include because it had orders of magnitude less eukaryotic species than the other tools (only 122 eukaryotic species as of release mpa_v30). The third approach uses a database of single copy orthologs that are found in most eukaryotes. The only tool that we found using this approach is EukDetect.

Workflow and implementation

The workflow of taxaTarget is shown in Figure 1. It is written in Python3 (Python ≥ 3.6) and executed via command line. It utilizes a custom database of eukaryotic marker genes, the trained classifiers, and the NCBI taxonomy. It accepts as input raw metagenomic sequencing reads in fasta/fastq format. The reads are first mapped to the database with Kaiju (v1.8.0) [33] to quickly identify reads that might be sequenced from eukaryotic marker genes. The binned reads are then more sensitively aligned with Diamond (v2.0.11) [34] to the marker genes. The Diamond output is then provided as input to taxon-specific classifiers trained on a curated database of eukaryotic sequences. Finally, the results are aggregated and the taxonomic profile of the metagenomic sample is generated.

Building the marker gene database

The database of taxaTarget was built from protein sequences extracted from the UniRef100 database of UniProt (release-2021_02) [35] to maximize the breadth of species detectable. The UniRef100 database was used, instead of UniProt, because it removes identical sequences and assigns the lowest (most recent) common ancestor to the representative sequences in the database. The $\sim 2.0 \times 10^8$ protein sequences in the UniRef100 database were downloaded on April 15, 2021. Approximately 6.4×10^7 eukaryotic proteins were extracted and assigned a taxonomic lineage using the taxonomic identifiers provided by UniRef100 and the NCBI taxonomy. Eukaryotic proteins were excluded if the taxonomic lineage included the keywords “unclassified eukaryotes”, “environmental samples”, or if the base taxonomic label was at a broader taxonomic rank than species. Eukaryotic marker genes were identified and extracted if they aligned with Diamond to the 255 eukaryotic Benchmarking Universal Single-Copy Orthologs (BUSCO) marker genes (version 5) [26] with an alignment matching or exceeding the BUSCO cutoffs for sequence length and bit score.

Training the taxonomic classifiers

The classifiers of taxaTarget are a set of region-specific functions that use thresholds for family, genus, and species, to classify metagenomic reads. The thresholds for species, genus, and family ($S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$, respectively) are learned for each non-overlapping 20 amino acid window, W , of each marker gene, M , in the taxaTarget database via training. As an example, the notation for the species threshold of

the 140 to 160 amino acid window of marker gene UniRef100_C0H4X5 (its UniRef100 identifier) would be $S_{(UniRef100_C0H4X5, 140)}$. The thresholds learned during training for each window provide the basis for the taxaTarget classifiers.

To provide an overview, the training process began by decomposing the UniRef100 database into a set of reads. Each read was aligned to the taxaTarget database, yielding one or more alignment scores depending on how many marker genes it aligned to. Per alignment, each read was assigned a label, either “species”, “genus”, “family”, or “negative”, depending on its taxonomic relationship to the marker gene. A read that aligned to a marker gene from a broader rank than family was considered a negative. The set of labelled reads that mapped within the window of a marker gene were used to determine the thresholds for that window.

The alignment score used for training and classification was the mean bit score—the bit score of a read alignment divided by the alignment length. The mean bit score was used because there is an approximately linear relationship between the alignment bit score and alignment length [31], i.e., reads of different lengths, sequenced from the same gene, will align back to the marker gene with different bit scores, but similar mean bit scores. Therefore, training on the mean bit score allows for standardized classification of reads of different lengths.

In detail, training began by aligning all UniRef100 reads (reads are all 70 amino acid long k-mers, with step size 10, extracted from UniRef100), to the marker genes using Diamond BLASTP. The procedure for identifying the species, genus, and family thresholds ($S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$, respectively) for each window, W , of each marker gene, M , in the taxaTarget database is as follows. To provide a specific example for ease of understanding, let us say we are trying to learn $S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$ for the 140 to 160 amino acid window of marker gene UniRef100_C0H4X5, i.e., $S_{(UniRef100_C0H4X5, 140)}$, $G_{(UniRef100_C0H4X5, 140)}$, $F_{(UniRef100_C0H4X5, 140)}$.

1. If a read alignment to UniRef100_C0H4X5 started or ended within the 140 to 160 amino acid window (extending towards the end of the gene) with a minimum bit score of 60 and alignment length of at least 30 amino acids,
 - a. The mean bit score of the read alignment was extracted
 - b. A taxonomic label (“species”, “genus”, “family”, or “negative”) was assigned depending on the taxonomic relationship of the read to the marker gene (*Plasmodium falciparum* for UniRef100_C0H4X5)
2. $S_{(UniRef100_C0H4X5, 140)}$, $G_{(UniRef100_C0H4X5, 140)}$, $F_{(UniRef100_C0H4X5, 140)}$ were set to the minimum mean bit score of the reads mapped and corresponding to that taxonomic rank
3. Thresholds were adjusted or removed based on the following conditions
 - a. A threshold was removed if there was no training data corresponding to its taxonomic rank

- b. If the mean bit score of a broader taxonomic rank was higher than the threshold of a more specific taxonomic rank, that threshold was either adjusted to that mean bit score or removed if the mean bit score was higher than the threshold of the next most specific rank
- c. The window was only kept in taxaTarget for classification if,

- i. $S_{(UniRef100_C0H4X5, 140)}$ had a value (in some cases reads did not map to a window)

- ii. $S_{(UniRef100_C0H4X5, 140)} > \{G_{(UniRef100_C0H4X5, 140)}, F_{(UniRef100_C0H4X5, 140)}, \max(\text{negatives})\}$

4. The process was repeated for every window of every marker gene in the database.

At the end of the training process, the values for $S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$ were defined for each window, for each marker gene, subject to the filtering parameters described above. This provided the basis for the classifiers used by taxaTarget for classification.

Taxonomic classification of metagenomic reads

Classifying individual reads

Taxonomic classification with taxaTarget begins by aligning metagenomic reads to the taxaTarget database with Diamond BLASTX, yielding alignment scores. The highest mean bit score alignment(s) of individual reads are used for classification. If a read aligns with 100% identity to multiple marker genes from different species, the read is classified at the lowest common ancestor of the respective hits. Otherwise, individual reads are classified as follows:

1. The alignment of a read to a marker gene must be at least 30 amino acids long (gaps are counted)
2. A read alignment must either start or end in the window of a classifier and extend toward the end of the marker gene. If a read alignment overlaps multiple windows, the window used for classification is $\text{window} = \text{floor}(\min(\text{start}, \text{end}) / 20)$. If a classifier does not exist for that window, the read is not classified.
3. The classifier for a window consists of a function that uses the learned classification thresholds for species, genus, and family, $S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$, respectively, where M is the specific marker gene the read aligned to and W is the window.
4. The mean bit score of the read alignment is provided as input to the classifier
5. Classification
 - a. If the mean bit score $\geq S_{(M, W)}$, classify the read at the species level
 - b. If the mean bit score $\geq G_{(M, W)}$, classify the read at the genus level
 - c. If the mean bit score $\geq F_{(M, W)}$, classify the read at the family level
 - d. Else, the read is not classified

Generating the taxonomic profile of the metagenomic sample

Once the individual reads have been taxonomically classified, taxaTarget uses several functions to analyze the pattern of mapped reads for each taxon and to filter out false positives from the taxonomic profile. The sources of false positives were identified through experiments or reported in the literature previously [22]. Each parameter has a default setting that can also be set by the user.

1. Thresholds determined from training data that lacked examples from broader taxonomic ranks or negatives are frequently overly permissive leading to false positive classifications. To address this factor, taxaTarget relies on “padding,” a user tunable parameter, to adjust the thresholds higher (requiring a higher mean bit score in order to classify a sequence). The default padding added (0.5 by default) was determined via parameter sweep analyses detailed later in the Methods. The following padding strategy is used:
 - a. By default, $T_{\text{padding}} = T + 0.5(I - T)$, where T is the threshold (options are $S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$) before padding, I is the mean bit score of the identity alignment for the window, and T_{padding} is the threshold value after padding is added
 - b. Padding is added using the following series of rules. Once a rule is met, the subsequent rules are not applied. During training, if no reads aligned to a marker gene,
 - i. At a rank broader than family, then padding is added to $S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$
 - ii. At a rank broader than genus, then padding is added to $S_{(M, W)}$ and $G_{(M, W)}$
 - iii. At a rank broader than species, then padding is added to $S_{(M, W)}$
2. As reported in [22] and supported by our observations, false positives frequently entail just one or two reads mapping to a single marker gene of an organism. To minimize such errors, taxaTarget requires that at least 3 reads be classified per taxonomic unit, and that reads map to at least 3 marker genes within a taxonomic unit for it to be reported in the output.
3. If sequencing samples the genomes in a metagenome randomly, then the reads should approximately cover a genome uniformly. False positives are often characterized by many reads mapping to a small subset of the marker genes from a particular taxon, perhaps due to a conserved domain. taxaTarget relies on outlier detection to filter out marker genes that have more mapped reads than expected for each taxon. The cutoff for the maximum expected number of mapped reads for a marker gene is approximated as follows:
 - a. Let $X = \{x_1, x_2, \dots, x_N\}$, where X is a vector of counts and x_i is a single draw, sampled with replacement, from the set of marker genes, $\{m_1, m_2, \dots, m_K\}$, where N is the total number of reads mapped to the marker genes, and K is the total number of marker genes for a taxon. $C = \text{mode}(X) \times 2$.
 - b. Any marker gene with more mapped reads than the mean of C after 100 experiments is excluded for the taxon being considered.

4. A source of error, also identified in [22], is that more species of a genus get detected than are present in the sample. Sometimes, this is due to the species lacking a complete set of marker genes in the taxaTarget database. To address such situations, taxaTarget examines the pattern of read mappings to the marker genes from the equivalent species using an extension of the strategy reported in [22]. First, a primary species is defined as the species with the most mapped reads that passes all previous filters. The primary species is assumed present, and the other species are compared to it to determine if they are present as well or false positives. For demonstration purposes, let us consider two scenarios (that taxaTarget handles) where two species, A and B, within the same genus have mapped reads, where A is the primary species.

- a. If the reads of species A and B map to non-overlapping sets of marker genes, the read counts are summed and the taxonomic profile reports both species as a single result indicating there is not enough information to determine whether either or both are present.
- b. If the reads of species A and B map to partially overlapping sets of marker genes, species B is considered a false positive if the ratio of reads mapped (normalized by the number of marker genes) to shared genes versus not-shared genes is less than 1, with a tolerance of 10%.

Simulated data sets

The simulated datasets were created using ART (Mount Rainier version) [36] (using default parameters except for, Illumina HiSeq 2500, single-end, 150 bp, 1X depth of coverage) for the genomes described in Table 1.

For the *non-microeukaryote* dataset, the GenBank assembly report (available at https://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/assembly_summary_genbank.txt) was parsed with a custom Python script to download one assembly per genus for each of the following phylogroups: metazoa (animals), embryophyta (plants), archaea, bacteria, and viruses. Per genus, the assemblies were first prioritized by assembly level, i.e., complete genome, chromosome, scaffold, contig. Then an assembly was randomly selected from the highest assembly level available for that genus. For analysis, simulated reads were combined into metagenomes by phylogroup and randomly down sampled to approximately 10 million reads. All reads simulated from the archaea and virus assemblies were used as they were already fewer than 10 million.

Assemblies from 54 species for the *microeukaryote_in_DB* dataset were downloaded from the NCBI assembly database after searching for “eukaryota” and filtering for “Latest RefSeq”, “Protists”, “Fungi”, “Complete genome”, “Chromosome”, “Representative.” Here, we only used high quality assemblies to reduce the likelihood of contaminating sequences in the assemblies which could affect the classification results. The assemblies were further filtered to only include species that had sequences in the databases of each tool tested (taxaTarget, EukDetect, mTAGs, and Metaxa2). The read sets were analyzed individually as well as a metagenome mixture where all the reads had been combined into a single fastq file.

The assemblies for the 83 species in the *microeukaryote_not_in_DB* were downloaded using the GenBank assembly report and a custom Python script that found species without sequences in any of the databases used by mTAGs, Metaxa2, EukDetect, and taxaTarget, but that did have sequences from the same genus in these databases. Each read set was analyzed individually.

Analyses using simulated data

Parameter sweep analysis for adjusting classification thresholds

As mentioned earlier, thresholds learned from training data that lacked examples from broader taxonomic ranks frequently allowed false positive classifications. To address this, taxaTarget relies on “padding” these thresholds, a user tunable parameter, to adjust the thresholds higher. The default amount of padding to add to thresholds was assessed via a parameter sweep analysis. The values tested for padding included (0.0, 0.1, 0.2, ..., 1.0). The experiment involved three simulated datasets which were useful for assessing different performance metrics of taxaTarget.

First, the *non-microeukaryote* dataset was used to assess the false positive rate of taxaTarget—both with and without applying outlier detection (see above). The false positive rate was calculated as the proportion of reads from non-protist/non-fungi phylogroups (i.e., animals, archaea, bacteria, plants and viruses) that were misclassified as a protist or fungus.

Second, the metagenome of the *microeukaryote_in_DB* dataset was used to assess the relationship between the precision and the sensitivity of taxaTarget for species with sequences in the database. The precision was calculated using the number of true positive and false positive reads classified at the species level. The sensitivity of taxaTarget was calculated using the number of true positive and false negative species detected in the metagenome.

Third, the *microeukaryote_not_in_DB* dataset was used to assess the relationship between the precision and sensitivity of taxaTarget for species with no sequences in the database, but that have relatives from the same genus within the database. The precision and sensitivity were calculated as before except at the genus level.

Limit of detection analysis

Simulated read sets, meant to mimic metagenomic samples where protist and fungal species occur at low abundance, were created as a control to test the limit of detection of the tools. The reads were simulated from the GenBank assemblies of three protist—*Blastocystis* subtype 1 (GCA_001651215.1), *Plasmodium falciparum* (GCA_000002765.2), and *Phytophthora sojae* (GCA_000149755.2)—and one fungal—*Aspergillus niger* (GCA_000002855.2)—species representing a range genome sizes (1.6×10^7 bp to 8.3×10^7 bp). Six read sets were created from each assembly, composed of 1, 10, 100, 1×10^3 , 1×10^4 , and 1×10^5 reads using a custom Python script to randomly select from the ART simulated reads. Every tool

tested had reference sequences from these species in their database. Each read set was classified separately with the tools and the number of reads correctly classified at the genus level were recorded.

Testing the false positive rate of the tools for non-protist/non-fungi species

The *non-microeukaryote* dataset (Table 1) was used to test the false positive rate of the tools. The false positive rate was calculated as the proportion of reads from non-protist/non-fungi phylogroups (i.e., animals, archaea, bacteria, plants and viruses) that were misclassified as a protist or fungi. The mean false positive rate for all phylogroups was normalized by 10 million to derive the mean number of false positive reads per 10 million reads classified.

Measuring the precision and sensitivity of the tools for protist and fungal species

The *microeukaryote_in_DB* dataset (Table 1) was used to measure the precision of the tools when classifying protist and fungal species with reference sequences in the database. The tools were first applied to each read set individually to provide a baseline for the precision of the tools and then to the combined read set to assess the precision of the taxonomic profile when multiple species were present (sometimes multiple species from the same genus). The precision was calculated separately at the taxonomic levels of family, genus and species. The sensitivity of the tools was defined at the family, genus and species levels as the proportion of the total number species in the dataset where at least one read was classified correctly at the respective taxonomic levels.

The *microeukaryote_not_in_DB* dataset (Table 1) was used to measure the precision of the tools when classifying protist and fungal species with no reference sequences (although there were reference sequences from other species within the corresponding genera) for the species in the databases used by the tools. The precision and sensitivity of the tools were measured as mentioned in the previous paragraph except only at the taxonomic levels of family and genus as the respective species are not represented in the reference collection.

Analysis of real metagenomic data

The accuracy of the tools for detecting known protist species in a microbiome was quantified using 57 Cameroonian shotgun metagenomic gut samples [37]. Polymerase chain reaction (PCR) had previously been performed on the samples to detect *Blastocystis* and microscopy to detect *Entamoeba*. Additionally, the previous study had identified several target protist species by mapping the metagenomic reads to a database of their genomes or 18S rRNA sequences [37]. We summarized and compared the taxonomic profiles of each tool with the findings of the original study. For *Blastocystis*, the PCR results were used as a “gold standard” for measuring the accuracy of the tools for detecting this parasite.

Results

The taxaTarget database

The reference database of taxaTarget is composed of 877,724 UniRef100 proteins which were identified as orthologs of the 255 BUSCO eukaryotic marker genes. The database sequences originate from 1,319 families, 2,564 genera and 4,232 species. As seen in Figure 2B, many species had few marker genes (e.g., 1,152 species only had one marker gene identified); however, 2,619 species had over 50 marker genes and the median number of marker genes per species was 152.

Statistics about the trained classifiers of taxaTarget

Classifiers were trained for all non-overlapping 20 amino acid regions (22.7×10^7 potential regions) per protein in the database. About $\frac{1}{3}$ of the regions (7.4×10^6) were excluded after training because the highest mean bit score alignments for family, genus, and species did not follow the pattern of species > genus > family (Figure 2A). For the remaining $\sim \frac{2}{3}$ of regions, training data was often sparse i.e., 22.7%, 64.2%, and 68.1% of the regions lacked negative-labeled, family-labeled, or genus-labeled data, respectively. The number of region-specific classifiers varied by orders of magnitude between marker genes, from 2.8×10^3 to 1.5×10^6 with a median of 4.3×10^4 . There was also great variability per species, from 1 (e.g., *Phytophthora andina*) to 1.6×10^5 (e.g., *Aureobasidium pullulans*) with a median of 2.1×10^3 region-specific classifiers per species. Furthermore, there were cases where the mean bit score showed high variability across windows of a marker gene within a single rank; the maximum mean bit score range observed was 1.94 across the windows of the Brix domain-containing protein of *Hortaea werneckii* (A0A3M7GIU8).

Parameter sweep analyses: Padding thresholds trained with sparse data

The first parameter sweep analysis assessed how threshold padding, with and without outlier detection, affected the number of non-microeukaryote reads falsely classified as microeukaryotic (protist or fungal) species (Supplementary Figure 1A) by taxaTarget. Outlier detection refers to whether taxaTarget is set to remove marker genes with more mapped reads than expected (see Methods).

Irrespective of padding, taxaTarget always performed better when using outlier detection, especially at low threshold padding, i.e., 0.2 or less (Supplementary Figure 1A). Without outlier detection, the number of false positive reads per 10 million reads classified decreased from 179 to 7 as the threshold padding increased from 0 to 1. With outlier detection, the same value over the same range decreased from 26 to 4. From 0.1 to 0.2 padding there was an unexpected increase in the number of false positive reads classified as *embryophyta*. This was caused by several species where the pattern of mapped reads to marker genes was not uniformly distributed at 0.1 padding—resulting in them being filtered out of the taxonomic profile by taxaTarget. However, at 0.2 padding the number of reads mapped to marker genes was greatly reduced and approximated uniform coverage—resulting in them being wrongly included in the taxonomic profile. The phylogroups that contributed to the most false positives, irrespective of padding

level, were the bacteria and *embryophyta*. For bacteria, misclassified reads were most frequently classified as fungi, *chlorophyta*, and *choanoflagellata*. For example, at 0.5 padding there were 15 misclassified reads, of which 7, 5, and 4 were assigned to fungi, *chlorophyta*, and *euglenozoa*, respectively. For *embryophyta*, misclassified reads were most frequently classified as *chlorophyta* and *charophyceae*—green algae. For example, at 0.5 padding there were 23 misclassified reads, 20 were assigned to *chlorophyta* (the other 3 were assigned to *stramenopiles*). The viruses, archaea, and *metazoa* did not produce false positives, only did so at a low rate (irrespective of threshold padding), or only at a padding of 0.1 or lower, respectively.

The second analysis used the *microeukaryote_in_DB* metagenome to assess the effect of padding thresholds on the precision and sensitivity of taxaTarget at the species level for species with sequences in the reference database (Supplementary Figure 1B). The sensitivity was 1.0 regardless of padding, i.e., all species were detected at the species level in the metagenome. With respect to precision, as padding increased from 0 to 1 there was a slight increase from 0.951 to 0.981, respectively.

The third analysis used the *microeukaryote_not_in_DB* dataset to assess the effect of padding thresholds on the precision and sensitivity of taxaTarget for species without sequences in the database (Supplementary Figure 1C). The precision was assessed at the genus level. Similarly, the sensitivity was calculated as the proportion of species detected at the genus level. Overall, the sensitivity was 0.952 from 0 to 0.8 padding and 0.94 thereafter. The precision continuously increased from 0.814 to 0.916 from a padding of 0 to 1, respectively.

Together, the parameter sweep analyses led us to choose a threshold padding of 0.5 as the default setting for taxaTarget, but users can adjust this parameter.

Limit of detection analysis

The limit of detection for each tool was tested by classifying read sets simulated from genomes, where the number of reads varied by 6 orders of magnitude (1 , 10 , 100 , 1×10^3 , 1×10^4 , 1×10^5). The datasets were simulated from the genomes of two protist species (*Plasmodium falciparum* and *Phytophthora sojae*) and one fungal species (*Aspergillus niger*). The genomes ranged in size from 23, 34, and 83 Mbp for *Plasmodium falciparum*, *Aspergillus niger* and *Phytophthora sojae*, respectively. Each species had reference sequences in the database of each tool. The sensitivity of a tool was measured as the number of reads correctly classified at the genus level.

Amongst the tools, taxaTarget showed the greatest sensitivity (Figure 3), only requiring 100 reads to detect *Aspergillus* and 1,000 reads to detect *Phytophthora*, *Blastocystis*, and *Plasmodium*. EukDetect required 10,000 reads to detect *Aspergillus*, *Blastocystis*, and *Phytophthora* and 100,000 reads to detect *Plasmodium*. Metaxa2 and mTAGs showed the least sensitivity, unable to detect *Phytophthora* given 100,000 reads, and requiring 10,000 to 100,000 reads to detect *Aspergillus*, *Blastocystis*, and *Plasmodium*.

False positive rate when classifying simulated metagenomes composed of non-fungi/non-protist phylogroups

Microbiomes can be composed of hundreds or thousands of microbial species—usually only a small proportion are microeukaryotes (mostly protists and fungi) [17, 38]. The *non-microeukaryote* dataset was used to quantify the false positive rate of the tools when classifying non-protist/non-fungal phylogroups.

As seen in Table 2, EukDetect had the lowest overall false positive rate of zero. The mean number of false positive reads per 10 million reads classified for Metaxa2, taxaTarget, and mTAGs was 4, 8, and 15 respectively. For taxaTarget, the false positive rate was mostly driven by bacteria classified as *Choanoflagellata* and plants classified as green algae (*Charophyceae* and *Chlorellales*).

Classification performance for “known” fungal and protist species

The precision and sensitivity of the tools for classifying protist/fungal species, with representative sequences in their databases, was tested under two conditions using the *microeukaryote_in_DB* dataset. First, the classification performance of the tools was tested on simulated genomic read sets from 54 species to provide a baseline. Second, a 54 species metagenome (combining the genomic read sets into a single sample) was used to measure the precision of the tools on a complex, multispecies sample where 24 of the 54 species came from a genus with more than one species in the metagenome. We note that the results of mTAGs could not be compared with the other tools at the species level because mTAGs uses operational taxonomic units (OTU). OTUs are taxonomic groups defined by the clustering of marker sequences, which can evolve at different rates, and thus can have variable taxonomic resolution e.g., strain, species, subgenus.

When classifying genomic read sets (Table 3) all tools showed a precision of at least 95%, although taxaTarget often had the highest. However, only taxaTarget had 100% sensitivity i.e., correctly identified all species. EukDetect, had the second highest sensitivity (90.7%) only misclassifying 5 species as other species. In contrast, mTAGs and Metaxa2 had the lowest sensitivity, classifying orders of magnitude less reads and identifying the least genera. For example, Metaxa2 only identified 31.5% of the species and only classified 320 reads (out of a total of 7.5 million) at the species level.

When classifying the combined metagenome (Table 3), all tools still showed high precision ($\geq 95\%$). Metaxa2 and taxaTarget had the highest precision at the genus ($\sim 99\%$) and species ($\sim 97\%$) levels, but taxaTarget was still the only tool with 100% sensitivity. EukDetect showed a substantial loss in sensitivity (only detecting one species per genus when multiple species were present), detecting 68.5% of the species, in contrast to 90% when data from each species was analyzed separately. Like the previous experiment, mTAGs and Metaxa2 still showed the lowest sensitivity.

Classification performance for “unknown” fungal and protist species

Next, we tested if the tools could detect protist/fungal species that were not represented in their databases, but that had marker sequences from the same genus. We tested the tools using the *microeukaryote_not_in_DB* dataset which consisted of simulated genomic read sets from 83 species. The 83 species had no reference sequences in the databases of any tool, but had sequences from species related at the genus level (Table 4). We again note that the results of mTAGs could not be compared with the other tools at the species level (refer to the first paragraph of the previous section for a detailed explanation).

In terms of sensitivity, taxaTarget identified the correct family and genus for over 95% of the species. In contrast, all the other tools had a sensitivity of ~50% or less. EukDetect was the most precise tool, assigning the correct family and genus to 99.9% and 90.4%, respectively, of the reads it classified. taxaTarget showed the second highest precision, assigning 94.5% and 89.1% of the classified reads to the correct family and genus, respectively. In contrast, Metaxa2 and mTAGs had ~90% precision at the family level and ~50% at the genus level.

Detection of known genera in the Cameroonian gut metagenomes

The accuracy of the tools for detecting known protist species in a microbiome was quantified using 57 Cameroonian shotgun metagenomic gut samples (Table 5). PCR had previously been performed on the samples to detect *Blastocystis* and microscopy to detect *Entamoeba*. Additionally, several target protist species had been detected by mapping the metagenomic reads to a database of their genomes or to their 18S rRNA [37].

For *Blastocystis*, the PCR results were used as the “gold standard” for the analysis. The accuracy of the tools was assessed at the genus level because the databases of the tools did not contain all the *Blastocystis* species detected by PCR. Metaxa2 had the highest agreement with the PCR results, 96.5%, followed by mTAGs (94.7%), taxaTarget (89.5%) and EukDetect (82.5%). All the tools detected *Blastocystis* in one sample that was PCR negative. For the 5 PCR positive samples where taxaTarget did not detect *Blastocystis*, three samples only had one or two classified reads—taxa with fewer than 3 classified reads are removed from the taxonomic profile by taxaTarget. The other two samples had three and seven reads, respectively, but they mapped to fewer than three marker genes, the minimum number required by taxaTarget to be included in the taxonomic profile.

For *Entamoeba*, microscopy had previously been found less sensitive than mapping the metagenomic reads to reference genomes—33% and 63% of samples were found positive using microscopy and read mapping, respectively [37]. For our analysis, *Entamoeba* was identified in 54% (mTAGs), 37% (Metaxa2), 23% (taxaTarget), and 0% (EukDetect) of the samples, respectively. All *Entamoeba* positive samples were a subset of those identified with metagenomic read mapping in the previous study. However, *Entamoeba* was only identified in 84% (mTAGs), 69% (taxaTarget), and 67% (Metaxa2) of the *Entamoeba* positive samples determined via microscopy. Many of the samples where taxaTarget did not detect *Entamoeba* had been filtered from the taxonomic profile because too few reads mapped, or the reads mapped to too

few marker genes. Before filtering, taxaTarget had identified *Entamoeba* in 95% of microscopy positive and 89% of read mapping positive samples.

Compared to the previous analysis [37], none of the tools detected *Endolimax* or *Enteromonas* at the genus level. However, despite not having database sequences from *Enteromonas*, taxaTarget still detected Hexamitidae (the family of *Enteromonas*) in 3 of 4 samples where *Enteromonas* had been identified previously. Further, the previous study had also identified *Giardia* in two samples—mTAGs detected *Giardia* in both whereas taxaTarget either detected *Giardia* or Hexamitidae (the family of *Giardia*).

Comparison of the taxonomic profiles for the 57 Cameroon gut samples

The total diversity of microeukaryotes identified in the Cameroon human gut metagenomes was compared for the four tools (Figure 4). Overall, 59 genera were detected—EukDetect, taxaTarget, Metaxa2, and mTAGs detected 21%, 35%, 39%, and 46% of the genera, respectively. Amongst these genera, only *Blastocystis* was detected by all tools—also had the highest occurrence in the samples. However, taxaTarget had the greatest overlap with the other tools, sharing 9 genera with EukDetect, 6 with mTAGs, and 3 with Metaxa2. The 3 genera taxaTarget did not share with EukDetect were either not in the database of taxaTarget (*Kodamaea*) or were not reported in the results for having too few marker genes in the database (*Volvariella*) or too few mapped reads (*Bactrocera*). For the 18 genera not shared with Metaxa2, 10 were not in the database of taxaTarget and 6 were not reported in the results for having too few marker genes in the database, too few mapped reads, or too few marker genes with mapped reads. Further, although *Aplanochytrium* was not detected at the genus level, its family was detected. Only the genus *Amphora* showed no signs of detection at the family or genus levels despite having sufficient marker genes in the taxaTarget database to be reported in the taxonomic report. For the remaining 18 genera not shared with mTAGs, mTAGs had detected each in a single sample by a single read. Six of these genera were not in the taxaTarget database, five were in the database but were not detected, two were detected at the family level, and five were not reported because too few reads mapped.

Consistent with other studies of the human gut eukaryome [38], no genus was found in all samples and most genera were rare. For perspective, there were only three (EukDetect), four (mTAGs), and five (Metaxa2 and taxaTarget) genera identified in five or more samples. For these genera, taxaTarget contained all detected by EukDetect and all but one by Metaxa2, *Korotnevella*, and by mTAGs, *Hesseltinella*.

Of the 20 genera detected by taxaTarget in the 57 Cameroonian gut samples, 16 had been observed in the human gut microbiome before (Table 6). Of the remaining four, *Chaetoceros*, *Perkinsus*, and *Wallemia* potentially could have been transiently introduced to the gut microbiome if a person consumed foods that carried them. Amongst the genera identified by taxaTarget were parasites with clinical relevance to humans such as *Trichinella*, *Schistosoma*, *Strongyloides*, *Digenea*, *Necator*, *Trypanosoma*, and *Giardia*.

Computational resource usage of the tools

The runtime and maximum random access memory (RAM) usage (Figure 5) of the tools were recorded for the *non-microeukaryote* and *microeukaryote_in_DB* simulated metagenomic datasets as well as one of the Cameroonian gut metagenomes—a total of ~91 million reads. All tools were run using 12 CPUs on a 256 gigabytes RAM node (Intel® Xeon® central processing unit, E5-2650, version 3, at 2.3 gigahertz). In terms of runtimes, mTAGs had the shortest (85 minutes) followed by taxaTarget and EukDetect (102 minutes). EukDetect (1.7 GB) and taxaTarget (2 GB) had the smallest maximum memory requirement followed by mTAGs (3.2 GB). In contrast, although Metaxa2 did not have a substantially longer runtime, 130 minutes, its memory usage ranged widely from a minimum of 46 MB to a maximum of 12 GB depending on the dataset being analyzed.

Discussion

Metagenomics is a powerful tool for exploring the complex microbial communities of bacteria, archaea, viruses, and eukaryotes that compose microbiomes. However, despite the many tools available, we are still not able to comprehensively assess the taxonomic composition of metagenomes. Here we report an important step forward for the identification of microeukaryotic species in metagenomic data by introducing taxaTarget, a fast and computationally efficient supervised learning method that learned region-specific classification thresholds from the UniRef100 database. Our results show this approach can match and often outperform other marker-gene-based tools in terms of precision and sensitivity. These gains applied for species with sequences in the database and for species which only have close relatives in the database—extending the predictive power of the scarce sequence data available for microeukaryotes. Although the use of a protein database for taxaTarget could be expected to have high sensitivity for divergent sequences, due to the redundancy of the genetic code, it was not expected to be more precise than tools using nucleotide databases (indicating that higher precision might be attained with a nucleotide database).

Scarce reference sequences for microeukaryotes present many challenges

Perhaps the greatest challenge for developing taxaTarget was the scarcity of reference sequence data for microeukaryotes versus their estimated species diversity [39]. It is imperative that more genomes/transcriptomes/proteomes are sequenced for eukaryotic species to increase the breadth of detectable taxa in metagenomes. As seen in the analysis of the Cameroonian gut metagenomes, although taxaTarget identified more microeukaryotes than the previous study of the dataset, including clinically relevant parasites, the biased representation of species in the databases of the tools acutely affected the results. EukDetect and taxaTarget—which use a similar marker gene database—detected an almost non-overlapping set of genera compared to Metaxa2 and mTAGs, which use databases of ribosomal subunits and COI genes. This result is somewhat expected given that the databases of

Metaxa2 and mTAGs contain nearly an order of magnitude more taxa than the databases of EukDetect and taxaTarget. However, what is unexpected is that Metaxa2 and mTAGs identified a total of 44 genera and yet only shared 4 despite using very similar databases. This calls into question the validity of many of these detections and the reproducibility of results when using these markers for taxonomic classification of metagenomic data.

Scarce data is especially problematic for supervised learning because the accuracy of classification is often dependent upon the number and diversity of labelled training data. For taxaTarget, the classification thresholds of many regions were learned using five or fewer data points, preventing the use of more sophisticated machine learning approaches. Some regions lacked some or all the thresholds because there was no training data. Together, this prevented the use of neighboring windows to impute missing thresholds. However, as more data becomes available, future work should test imputation methods (a common solution for handling missing data in supervised learning [40]) as they might increase sensitivity/precision by including more regions for classification.

Solutions for handling scarce reference sequences

A highly effective solution, implemented in taxaTarget, for handling classification errors due to scarce data was to pad thresholds that lacked training data at broader taxonomic ranks. However, even at high threshold padding, false positives from non-microeukaryotes persisted, highlighting the need for more comprehensive training data. A persistent source of classification errors were small regions of marker genes with high sequence similarity in non-microeukaryotes. Provided sufficient depth of metagenomic sequencing, such regions might be simply filtered using outlier detection [41]; however, we often observed that each microeukaryotic taxa identified in a metagenome had less than 1000 mapped reads—far less than 1X depth of coverage. The simple and effective solution developed for taxaTarget used a variant of the “urn problem” to detect and remove outlier marker genes. The marker genes of a taxa were treated like marbles with different colors in an urn and the number of reads mapped to the marker genes were treated as draws, with replacement, from the urn. Averaged over 100 simulations, the count of the marker gene drawn most often provided a useful baseline for the highest expected variant.

Another source of error caused by scarce data, (observed here and previously [22]) are species that have not had their full set of marker genes sequenced. For example, if *Entamoeba histolytica* has a full complement of marker genes in the database, but *Entamoeba dispar* does not, metagenomic reads, corresponding to the missing genes of *E. dispar*, could map to the orthologs in *E. histolytica* and thus get misclassified. EukDetect partially addressed this problem by filtering out species where more reads mapped to unshared versus shared marker genes with another species in the same genus. taxaTarget implemented a modified version of this solution and found it effective, although some errors persisted. EukDetect applied an additional filter (taxaTarget does not use), to further address this source of error, which removes species if the global percent identity of aligned reads across genes shared with the primary species is less than that for the primary species. It was acknowledged in [22] that this might

result in only detecting the primary species of a genus, but this bias was observed for EukDetect even when using a simulated metagenome where each species had 1X depth of coverage. Future work should explore the adaptation of statistical methods for assessing ambiguity in taxonomic classification, learned from amplicon sequencing [30].

Insights from training and improvements

A major benefit of using a supervised learning approach was that training provided granular information about the discriminatory power of each region of the database sequences (Metaxa2, mTAGs, and EukDetect do not provide), which sometimes varied widely for adjacent regions (this information is provided for download with taxaTarget and might be useful for the study of sequence evolution). Where EukDetect removed 41 of the 255 eukaryotic BUSCO marker gene families from its database for regions conserved in bacteria, taxaTarget was able to retain all 255, by only excluding the regions that lacked discriminatory power.

Previously, the most expensive step for creating a tool like taxaTarget was the training process. To classify metagenomic reads of variable length, MetaPhyler and ROcker were trained with a minimum of three read sets of different length [31, 32]. However, we found that the longest read set was sufficient for training because the mean bit scores of the alignments were similar to those of the shorter read lengths. This resulted in a threefold reduction in training time. A potential future research area is exploring the replacement of the read mapping step of training with directly using the pairwise alignments of the sequences in the database.

Future work to extend taxaTarget

Future work should test the methods of taxaTarget when extended to the marker genes of other phylogroups such as bacteria, archaea, and viruses. Furthermore, the methods should be tested when extended to entire sequence databases—which might improve sensitivity and specificity. Ideally, given sufficient data, training could identify which regions of a genome or proteome are useful for taxonomic classification. However, the use of sequences other than marker genes might not be feasible without the development of additional algorithms for handling scarce data and the resulting classification errors. Lastly, despite highlighting the performance gains of data-driven methods, taxaTarget still implemented multiple user-defined thresholds, which also might be optimized or replaced with data-driven methods.

Conclusions

Our results show that the taxonomic classification of microeukaryotes in metagenomic data can be improved by utilizing data-driven methods that learn classification thresholds from the structure of an input database. These gains applied for species with sequences in the database and for species which only have close relatives in the database. The scarcity of available reference sequences still presents

many challenges for the metagenomic taxonomic classification of microeukaryotes. More genomes/transcriptomes/proteomes need to be sequenced and methods for handling scarce data and the resulting classification errors need to be further explored. The implementation of these methods will help facilitate a more taxonomically comprehensive analysis of metagenomic data and expand our knowledge about roles and diversity of eukaryotes in microbial communities.

Declarations

Data availability

The source code and a tutorial for installing and running taxaTarget can be found on GitHub at <https://github.com/SethCommichaux/taxaTarget>. The taxaTarget database can be downloaded from <https://obj.umiacs.umd.edu/taxatarget/archive.tar.gz>.

All data used for analysis in this study are publicly available. The Cameroonian gut metagenomes can be found under NCBI BioProject PRJEB27005. The NCBI Assembly accessions for the *non-microeukaryote* dataset, *microeukaryote_in_DB*, and *microeukaryote_not_in_DB* datasets can be found in the supplementary materials.

Acknowledgements

Not applicable

Funding

MP was funded in part by the NIH, award R01-AI-100947. KJ was supported by the University of Maryland, Joint Institute for Food Safety and Applied Nutrition (JIFSAN) through cooperative agreement #5U01-FD001418, provided by the U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition (FDA, CFSAN).

Competing interests

The authors declare that they have no competing interests

Author's contributions

SC, AO, MP, and HR designed the study. SC, MP, HR, KJ, and HM designed the algorithms. SC analyzed the data and wrote the manuscript. SC, MP, HR, KJ, and AO interpreted the data. All authors read, provided feedback, and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Rights and permissions

Open source

References

1. Hawksworth DL, Lücking R: **Fungal Diversity Revisited: 2.2 to 3.8 Million Species.** *Microbiol Spectr* 2017, **5**(4).
2. Larsen BB, Miller EC, Rhodes MK, Wiens JJ: **Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life.** *The Quarterly Review of Biology* 2017, **92**(3):229-265.
3. Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, Bowser SS, Cepicka I, Decelle J, Dunthorn M *et al*: **CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms.** *PLoS Biology* 2012, **10**(11):e1001419.
4. Burki F, Roger AJ, Brown MW, Simpson AGB: **The New Tree of Eukaryotes.** *Trends in Ecology & Evolution* 2020, **35**(1):43-55.
5. Rousseaux C, Gregg W: **Interannual Variation in Phytoplankton Primary Production at A Global Scale.** *Remote Sensing* 2013, **6**(1):1-19.
6. Wong Z-W, Faulder K, Robinson JL: **Does Dientamoeba fragilis cause diarrhea? A systematic review.** *Parasitol Res* 2018, **117**(4):971-980.
7. Deng L, Wojciech L, Gascoigne NRJ, Peng G, Tan KSW: **New insights into the interactions between Blastocystis, the gut microbiota, and host immunity.** *PLoS Pathog* 2021, **17**(2):e1009253.
8. Hirt RP: **Mucosal microbial parasites/symbionts in health and disease: an integrative overview.** *Parasitology* 2019, **146**(9):1109-1115.
9. Campo J, Bass D, Keeling PJ: **The eukaryome: Diversity and role of microeukaryotic organisms associated with animal hosts.** *Functional Ecology* 2019.
10. Parfrey LW, Walters WA, Lauber CL, Clemente JC, Berg-Lyons D, Teiling C, Kodira C, Mohiuddin M, Brunelle J, Driscoll M *et al*: **Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity.** *Front Microbiol* 2014, **5**:298.

11. Even G, Lokmer A, Rodrigues J, Audebert C, Viscogliosi E, Ségurel L, Chabé M: **Changes in the Human Gut Microbiota Associated With Colonization by sp. and spp. in Non-Industrialized Populations.** *Front Cell Infect Microbiol* 2021, **11**:533528.
12. New FN, Brito IL: **What Is Metagenomics Teaching Us, and What Is Missed?** *Annu Rev Microbiol* 2020, **74**:117-135.
13. Laforest-Lapointe I, Arrieta M-C: **Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies.** *mSystems* 2018, **3**(2).
14. Hooks KB, O'Malley MA: **Contrasting Strategies: Human Eukaryotic Versus Bacterial Microbiome Research.** *J Eukaryot Microbiol* 2019.
15. Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB: **Protists are microbes too: a perspective.** *The ISME Journal* 2009, **3**(1):4-12.
16. Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK: **Sequencing our way towards understanding global eukaryotic biodiversity.** *Trends Ecol Evol* 2012, **27**(4):233-243.
17. Jacquiod S, Stenbæk J, Santos SS, Winding A, Sørensen SJ, Priemé A: **Metagenomes provide valuable comparative information on soil microeukaryotes.** *Research in Microbiology* 2016, **167**(5):436-450.
18. Saary P, Mitchell AL, Finn RD: **Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC.** *Genome Biol* 2020, **21**(1):244.
19. West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF: **Genome-reconstruction for eukaryotes from complex natural microbial communities.** *Genome Res* 2018, **28**(4):569-580.
20. Lu J, Salzberg SL: **Removing contaminants from databases of draft genomes.** *PLoS Comput Biol* 2018, **14**(6):e1006277.
21. Steinegger M, Salzberg SL: **Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank.** *Genome Biol* 2020, **21**(1):115.
22. Lind AL, Pollard KS: **Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing.** *Microbiome* 2021, **9**(1):58.
23. Nousias O, Montesanto F: **Metagenomic profiling of host-associated bacteria from 8 datasets of the red alga *Porphyra purpurea* with MetaPhlAn3.** *Mar Genomics* 2021:100866.
24. Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH: **METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data.** *Mol Ecol Resour* 2015, **15**(6):1403-1414.
25. Salazar G, Ruscheweyh H-J, Hildebrand F, Acinas SG, Sunagawa S: **mTAGs: taxonomic profiling using degenerate consensus reference sequences of ribosomal RNA genes.** *Bioinformatics* 2022, **38**(1):270-272.
26. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 2015, **31**(19):3210-3212.

27. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ: **The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages.** *Proc Natl Acad Sci U S A* 2009, **106**(18):7273-7280.
28. Marcelino VR, Clausen PTLC, Buchmann JP, Wille M, Iredell JR, Meyer W, Lund O, Sorrell TC, Holmes EC: **CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data.** *Genome Biol* 2020, **21**(1):103.
29. Nasko DJ, Koren S, Phillippy AM, Treangen TJ: **RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification.** *Genome Biology* 2018, **19**(1):165.
30. Shah N, Meisel JS, Pop M: **Embracing Ambiguity in the Taxonomic Classification of Microbiome Sequencing Data.** *Front Genet* 2019, **10**:1022.
31. Liu B, Gibbons T, Ghodsi M, Pop M: **MetaPhyler: Taxonomic profiling for metagenomic sequences.** *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2010.
32. Orellana LH, Rodriguez-R LM, Konstantinidis KT: **ROCKr: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores.** *Nucleic Acids Res* 2017, **45**(3):e14.
33. Menzel P, Ng KL, Krogh A: **Fast and sensitive taxonomic classification for metagenomics with Kaiju.** *Nat Commun* 2016, **7**:11257.
34. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND.** *Nature Methods* 2015, **12**(1):59-60.
35. Consortium TU, The UniProt C: **The Universal Protein Resource (UniProt).** *Nucleic Acids Research* 2007, **36**(Database):D190-D195.
36. Huang W, Li L, Myers JR, Marth GT: **ART: a next-generation sequencing read simulator.** *Bioinformatics* 2012, **28**(4):593-594.
37. Lokmer A, Cian A, Froment A, Gantois N, Viscogliosi E, Chabé M, Ségurel L: **Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels.** *PLoS One* 2019, **14**(2):e0211139.
38. Nash AK, Auchtung TA, Wong MC, Smith DP, Gesell JR, Ross MC, Stewart CJ, Metcalf GA, Muzny DM, Gibbs RA *et al*: **The gut mycobiome of the Human Microbiome Project healthy cohort.** *Microbiome* 2017, **5**(1):153.
39. Sibbald SJ, Archibald JM: **More protist genomes needed.** *Nat Ecol Evol* 2017, **1**(5):145.
40. Kang H: **The prevention and handling of the missing data.** *Korean Journal of Anesthesiology* 2013, **64**(5):402.
41. Deng C, Daley T, Calabrese P, Ren J, Smith AD: **Predicting the Number of Bases to Attain Sufficient Coverage in High-Throughput Sequencing Experiments.** *Journal of Computational Biology* 2019, **27**(7):1130-1143.
42. Nieves-Ramírez ME, Partida-Rodríguez O, Laforest-Lapointe I, Reynolds LA, Brown EM, Valdez-Salazar A, Morán-Silva P, Rojas-Velázquez L, Morien E, Parfrey LW *et al*: **Asymptomatic Intestinal Colonization**

- with Protist Blastocystis Is Strongly Associated with Distinct Microbiome Ecological Patterns.** *mSystems* 2018, **3**(3).
43. Iyer LR, Verma AK, Paul J, Bhattacharya A: **Phagocytosis of Gut Bacteria by Entamoeba histolytica.** *Frontiers in Cellular and Infection Microbiology* 2019, **9**.
 44. Spatz M, Richard ML: **Overview of the Potential Role of Malassezia in Gut Health and Disease.** *Frontiers in Cellular and Infection Microbiology* 2020, **10**.
 45. Hallen-Adams HE, Suhr MJ: **Fungi in the healthy human gastrointestinal tract.** *Virulence* 2017, **8**(3):352-358.
 46. Campa-Córdova AI, Luna-González A, Ascencio F, Cortés-Jacinto E, Cáceres-Martínez CJ: **Effects of chloramphenicol, erythromycin, and furazolidone on growth of Isochrysis galbana and Chaetoceros gracilis.** *Aquaculture* 2006, **260**(1-4):145-150.
 47. Dimidi E, Cox SR, Rossi M, Whelan K: **Fermented Foods: Definitions and Characteristics, Impact on the Gut Microbiota and Effects on Gastrointestinal Health and Disease.** *Nutrients* 2019, **11**(8).
 48. Afrin T, Murase K, Kounosu A, Hunt VL, Bligh M, Maeda Y, Hino A, Maruyama H, Tsai IJ, Kikuchi T: **Sequential Changes in the Host Gut Microbiota During Infection With the Intestinal Parasitic Nematode Strongyloides venezuelensis.** *Frontiers in Cellular and Infection Microbiology* 2019, **9**.
 49. Santiago-Rodriguez TM, Fornaciari G, Luciani S, Dowd SE, Toranzos GA, Marota I, Cano RJ: **Taxonomic and predicted metabolic profiles of the human gut microbiome in pre-Columbian mummies.** *FEMS Microbiol Ecol* 2016, **92**(11).
 50. Gouba N, Raoult D, Drancourt M: **Eukaryote culturomics of the gut reveals new species.** *PLoS One* 2014, **9**(9):e106994.
 51. Berrilli F, Di Cave D, Cavallero S, D'Amelio S: **Interactions between parasites and microbial communities in the human gut.** *Frontiers in Cellular and Infection Microbiology* 2012, **2**.
 52. Zajc J, Gunde-Cimerman N: **The Genus From Contamination of Food to Health Threat.** *Microorganisms* 2018, **6**(2).
 53. Wijayalath W, Majji S, Kleschenko Y, Pow-Sang L, Brumeanu TD, Villasante EF, Vasta GR, Fernández-Robledo J-A, Casares S: **Humanized HLA-DR4 Mice Fed with the Protozoan Pathogen of Oysters Perkinsus Marinus (Dermo) Do Not Develop Noticeable Pathology but Elicit Systemic Immunity.** *PLoS ONE* 2014, **9**(1):e87435.
 54. Dacks J: **Faculty Opinions recommendation of On the reversibility of parasitism: adaptation to a free-living lifestyle via gene acquisitions in the diplomonad Trepomonas sp. PC1.** *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature* 2017.
 55. Bucková B, Hurníková Z, Lauková A, Revajová V, Dvorožňáková E: **The anti-parasitic effect of probiotic bacteria via limiting the fecundity of Trichinella spiralis female adults.** *Helminthologia* 2018, **55**(2):102-111.

Tables

Table 1: The composition of the simulated metagenomic datasets

Dataset/Analysis	Phylogroup	Number of genera	Number of species/assemblies	Number of reads simulated
<i>non-microeukaryote</i>	Animals	383	383	10,022,536
	Archaea	139	139	2,511,659
	Bacteria	2,761	2,761	10,653,348
	Plants	77	77	10,140,188
	Viruses	2,171	2,171	464,121
<i>microeukaryote_in_DB</i>	Protists, Fungi (species represented in all tool databases)	39	54	7,536,554
<i>microeukaryote_not_in_DB</i>	Protists, Fungi (species not represented in database of any tool)	83	83	32,052,306
<i>Limit of detection analysis</i>	<i>Plasmodium falciparum</i>	1	1	100,000
	<i>Phytophthora sojae</i>	1	1	100,000
	<i>Aspergillus niger</i>	1	1	100,000
	<i>Blastocystis</i> subtype 1	1	1	100,000

Table 2: The false positive rate of the tools for the simulated metagenomes from non-protist/non-fungi phylogroups

Tool	Archaea	Bacteria	Plants	Animals	Viruses	Mean number of false positive reads per 10 million reads classified
EukDetect	0	0	0	0	0	0
Metaxa2	0	6.6×10^{-7}	5.9×10^{-7}	2.0×10^{-7}	0	4
taxaTarget	1.2×10^{-6}	1.5×10^{-6}	2.3×10^{-6}	0	0	8
mTAGs	4.0×10^{-7}	3.7×10^{-6}	8.9×10^{-7}	2.0×10^{-7}	0	15

Table 3: Precision of tools when classifying read sets (1X depth of coverage) simulated from the genomes of 54 protist/fungal species with reference sequences in the database of each tool. In the

experiment column, genome indicates the read sets were analyzed individually, whereas metagenome indicates the read sets were combined into a single sample that was analyzed.

Experiment	Tool	Precision at family level (%)	Precision at genus level (%)	Precision at species level (%)	Percentage of species detected (n = 54)	Reads classified at species level (n = 7.5x10 ⁶)
Genome	taxaTarget	99.0	98.9	97.3	100.0	35,950
	EukDetect	96.0	96.0	95.3	90.7	51,489
	Metaxa2	66.8*	97.9	97.4	31.5	320
	mTAGs	98.8	96.3	NA	NA	703 (OTU**)
Metagenome	taxaTarget	99.0	99.0	97.3	100.0	36,182
	EukDetect	96.2	96.2	95.0	68.5	44,823
	Metaxa2	76.3*	99.2	97.0	31.5	297
	mTAGs	97.5	94.9	NA	NA	713 (OTU**)

*Metaxa2 uses a custom and older taxonomic database which often does align with the current NCBI taxonomy at the family level. If the family label did not align with the current NCBI taxonomy it was counted as incorrect.

**Below genus, mTAGs classifies reads to OTUs and does not provide species level classifications

Table 4: Analysis of 83 single species simulated read sets. The 83 protist/fungal species had no reference sequences in the database of the tools, but had sequences from species related at the genus level.

Tool	Percentage of species detected at the family level (%)	Percentage of species detected at the genus level (%)	Precision family (%)	Precision genus (%)
taxaTarget	97.6	95.2	94.5	89.1
EukDetect	53.0	50.6	99.9	90.4
Metaxa2	27.7	16.9	90.3	41.6
mTAGs	47.0	36.1	88.1	50.8

Table 5: Detection of *Blastocystis* in 57 Cameroonian gut samples. PCR(+) indicates that PCR detected *Blastocystis* in the sample. Tool(+) indicates the tool detected *Blastocystis* in the metagenomic sample.

Tool	PCR(+) Tool(+)	PCR(-) Tool(-)	PCR(+) Tool(-)	PCR(-) Tool(+)	Accuracy (%)
Metaxa2	48	7	1	1	96.5
mTAGs	48	6	1	2	94.7
taxaTarget	44	7	5	1	89.5
EukDetect	40	7	9	1	82.5

Table 6: The 20 genera detected by taxaTarget in the 57 Cameroonian gut microbiome samples

Cameroon (prevalence in 57 samples)	Observed in human gut before	Relationships to humans
<i>Blastocystis</i> (77%)	yes	commensal/parasite stramenopile that can consume bacteria in gut [42]
<i>Entamoeba</i> (23%)	yes	commensal/parasitic amoeba that can consume bacteria in gut [43]
<i>Malassezia</i> (19%)	yes	broadly dispersed fungi, once thought to be restricted skin [44]
<i>Saccharomyces</i> (19%)	yes	broadly distributed yeast, some species are commonly found in food [45]
<i>Chaetoceros</i> (12%)	no	none, primarily marine diatoms; might transiently occur in human gut if a colonized shellfish is eaten [46]
<i>Pichia</i> (7%)	yes	yeast commonly found in food [47]
<i>Strongyloides</i> (5%)	yes	parasitic nematode [48]
<i>Trypanosoma</i> (4%)	yes	protist parasite [49]
<i>Phytophthora</i> (4%)	yes	plant pathogen [50]
<i>Aspergillus</i> (4%)	yes	diverse genus of mold consisting of several hundred species [45]
<i>Digenea</i> (4%)	yes	parasitic flatworm (fluke) [51]
<i>Schistosoma</i> (4%)	yes	parasitic flatworm (fluke) [51]
<i>Giardia</i> (2%)	yes	protist parasite [51]
<i>Wallemia</i> (2%)	no	fungi commonly found in food; can cause skin infections; might transiently occur in human gut if colonized food is consumed [52]
<i>Perkinsus</i> (2%)	no	molluscan parasite; might transiently occur in human gut if colonized molluscan is consumed [53]
<i>Fusarium</i> (2%)	yes	mycotoxin-producing fungi commonly [45] found on food
<i>Necator</i> (2%)	yes	parasitic nematode [51]
<i>Trepomonas</i> (2%)	no	free-living, environmental protist whose ancestor might have been parasitic [54]
<i>Brettanomyces</i> (2%)	yes	yeast commonly found on fruit [47]
<i>Trichinella</i> (2%)	yes	parasitic nematode that can affect microbial composition of

Figures

Figure 1

Workflow diagram for taxaTarget: A) Metagenomic reads are first mapped to the marker genes with Kaiju—a fast, k-mer approach. The subset of reads that mapped to the marker genes with Kaiju are then more sensitively aligned to the marker genes with Diamond. The mean bit score and start position of the Diamond read alignments are used for classification. B) An example marker gene that a read mapped to. Most windows (every non-overlapping 20 amino acid region) have classification thresholds (Species, Genus, and Family) learned from training, but some regions lacked training data and thus lack thresholds. C) Classification of a read for a window where threshold padding is applied. The start position of the read alignment is used to determine the window used for classification. During training, no negative labelled reads aligned to the window at a rank broader than genus and so the thresholds for Genus and Species get padded (i.e., adjusted upwards). Here, the read gets classified as the genus of the marker gene because the mean bit score of its alignment is above Genus with padding.

Figure 2

A) The pie chart shows the maximum taxonomic resolution of the regions learned by training. Nearly 1/3 of the 22.7×10^7 non-overlapping 20 amino acid regions (from 8.8×10^5 proteins) were not useful for species level classification because the highest bit score alignments were from broader taxonomic ranks.

B) The scatter plot shows the number of BUSCO marker genes (255 is a complete set) versus the number of region-specific classifiers for all 4,232 species in the taxaTarget database.

Figure 3

The limit of detection for each tool was tested using simulated data where the number of reads varied by 6 orders of magnitude (1, 10, 100, 1x10³, 1x10⁴, 1x10⁵). A read was considered correctly classified if the tool assigned it to the correct genus. The reads were simulated from 4 genomes which varied in length from ~16.6 Mbp to ~82.6 Mbp. The genomes came from 3 protist species (*Blastocystis* subtype 1, *Plasmodium falciparum* and *Phytophthora sojae*) and 1 fungal species (*Aspergillus niger*)—genome size in parentheses besides the species name.

Figure 4

Genera of microeukaryotes found by the tools in the 57 Cameroonian gut samples. A) In total, 59 genera were detected by the tools. None of the tools detected all the genera, but mTAGs detected 44%. B) taxaTarget had the greatest overlap with other tools for detected genera. Only 24% of the genera were detected by more than one tool—taxaTarget detected 86% of them. taxaTarget and EukDetect, which use a similar database of marker genes, shared 9 detected genera, the most between any tools. Metaxa2 and mTAGs, despite using a similar database of marker genes, only had 4 detected genera in common.

Figure 5

Computational efficiency of the tools when run on metagenomic data. The maximum random access memory usage (A) and runtimes (B) of the tools were recorded for the *non-microeukaryote* and *microeukaryote_in_DB* simulated metagenomic datasets as well as one of the Cameroonian gut metagenomes—a total of ~91 million reads.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterials.docx](#)
- [SupplementaryinformationdataNCBIIDs.xlsx](#)